УДК 519.6

ОСОБЕННОСТИ ПРОЦЕССА НАКОПЛЕНИЯ ПОГРЕШНОСТЕЙ ПРИ РЕШЕНИИ ЗАДАЧ ДЛЯ УРАВНЕНИЯ ДИФФУЗИИ КОНЕЧНОРАЗНОСТНЫМИ МЕТОДАМИ

В.П. Житников, Н.М. Шерыхалина, С.С. Поречный

Аннотация

Рассматривается смешанная задача для одномерного уравнения теплопроводности с несколькими вариантами начальных и краевых условий. Для решения применяются явная и неявная схемы. Для неявной схемы при решении системы уравнений используются методы прогонки и итераций. Для анализа погрешностей метода и округления применяется численная фильтрация конечной последовательности результатов, полученной для различных сеток с возрастающим числом узловых точек n. Кроме того, для исследования погрешности округления сравниваются результаты, полученные при нескольких длинах мантиссы машинного слова.

Результаты вычислительного эксперимента показали, что погрешность численного метода представляется суммой нескольких компонент степенного вида $c_j n^{-k}$ с целыми показателями ($k = k_j > 0$). Погрешность округления накапливается с увеличением числа узлов как cn^2. В отличие от методов численного дифференцирования и интегрирования функций эта зависимость от n носит детерминированный характер. Значение коэффициента c ограничено величиной, равной 10^{-M} (M – длина мантиссы), либо пороговой величиной погрешности при применении метода итераций. При изменении числа Куранта значение коэффициента c меняется хаотически.

Ключевые слова: погрешности численного решения, переход от случайной к детерминированной модели, модель источников погрешности, рациональные числа Куранта

1. Введение

Ранее было проведено исследование составляющих погрешностей различных численных методов с помощью фильтрации [1, 2, 3].

Фильтрация проводилась на основе априорной модели зависимости результата вычисления искомой величины z от параметра дискретизации n. Для многих вычислительных методов: численного дифференцирования, интегрирования и т. д. эта зависимость представляется в виде суммы нескольких слагаемых

$$z_n - z = c_1 n^{-k_1} + c_2 n^{-k_2} + \ldots + c_L n^{-k_L} + \Delta(n), \qquad (1)$$

где z_n – результат вычисления, полученный при значении параметра дискретизации (числа узлов сетки, числа слагаемых в сумме и т.п.), равном n; $k_1, \ldots, k_L$ – произвольные известные действительные числа ($k_1 < k_2 < \ldots < k_L$); c_j – неизвестные коэффициенты.

Составляющая $\Delta(n)$ может состоять из не вошедших в сумму слагаемых, остаточного члена, погрешности округления и других составляющих, порожденных несовершенством численного алгоритма и его программной реализации. Существенно то, что величина $\Delta(n)$ не имеет априорной оценки, и предполагается возможным возрастание этой величины при возрастании n.

Для некоторых численных методов компоненты зависимости погрешности от n имеют показательный вид [3, 4].

Исследование погрешностей различных численных методов: численного дифференцирования, интегрирования, решение различных задач для дифференциальных уравнений [1, 2] показало наличие нескольких компонент погрешности степенного вида (1). С помощью фильтрации удалось подавить эти регулярные компоненты и уменьшить погрешность искомых параметров. Подавление этих компонент позволило исследовать поведение составляющей $\Delta(n)$, главная часть которой состояла из погрешности округления, вызванной ограниченностью длины мантиссы машинного слова. В перечисленных методах зависимость $\Delta(n)$ имела хаотичный характер, и моделировалась случайной величиной с нулевым математическим ожиданием и среднеквадратичным отклонением, растущим как cn^k с ростом n (где c – величина близкая к 10^{-M}, M – длина мантиссы машинного слова, $k>0$ – показатель, зависящий от рассматриваемого метода).

При подробном исследовании результатов решения задач для уравнения теплопроводности выяснилось, что $\Delta(n)$ имеет не случайный характер, а возникают зависимости от параметров n, x и t, имеющие вполне детерминированный вид. Это будет показано ниже.

2. Постановка задачи

Рассматривается смешанная задача для одномерного уравнения диффузии (теплопроводности)

$$\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = 0, \quad (0 < x < 1,\ 0 < t \le T). \tag{2}$$

Начальное условие

$$u(x,0) = f(x), \quad (0 < x < 1). \tag{3}$$

Краевые условия

$$u(0,t) = \varphi_1(t),\ u(1,t) = \varphi_2(t), \quad (t \ge 0). \tag{4}$$

Рассматривается 3 варианта краевых и начальных условий

I)
$$f(x) = \sin(\pi x), \quad \varphi_1(t) = \varphi_2(t) = 0, \tag{5}$$

II)
$$f(x) = 0, \quad \varphi_1(t) = 1,\ \varphi_2(t) = 0, \tag{6}$$

III)
$$f(x) = 1 - x, \quad \frac{\partial u}{\partial x}(0,t) = \frac{\partial u}{\partial x}(1,t) = 0. \tag{7}$$

Точные решения:

– для варианта I
$$u(x,t) = e^{-\pi^2 t}\sin(\pi x), \tag{8}$$

– для варианта II
$$u(x,t) = -\frac{2}{\pi}\sum_{m=1}^{\infty}\frac{1}{m}e^{-m^2\pi^2 t}\sin m\pi x + 1 - x, \tag{9}$$

– для варианта III

$$u(x,t) = \frac{1}{2} + \frac{4}{\pi^2}\sum_{m=1}^{\infty}\frac{1}{(2m-1)^2}e^{-(2m-1)^2\pi^2 t}\cos(2m-1)\pi x. \tag{10}$$

Задачи с вариантами условий II и III моделируют эксперименты, проводимые с образцами пород, добытых при бурении скважин [5].

Задачи решаются методом конечных разностей. На плоскости (x,t) строится сетка с шагом h по переменной x ($x_i{=}ih$, $i{=}0,..,n$, $h{=}1/n$) и с шагом τ по переменной t ($t_j{=}j\tau$, $j{=}0,..,m$, $\tau{=}T/m$). Вводятся обозначения $u(x_i,t_j)=u_{i,j}$.

Обозначим $\lambda=\tau/h^2$ (число Куранта). Тогда разностные уравнения для неявной схемы запишутся в виде

$$-\lambda u_{i-1,j}+(1+2\lambda)u_{i,j}-\lambda u_{i+1,j}=u_{i,j-1}, i=1,..,n-1; j=1,..,m, \tag{11}$$

$$u_{i,0}=f(x_i), \quad u_{0,j}=\varphi_1(t_j), \quad u_{n,j}=\varphi_2(t_j).$$

Для явной схемы

$$u_{i,j+1}=\lambda u_{i-1,j}+(1-2\lambda)u_{i,j}+\lambda u_{i+1,j}, i=1,..,n-1; j=0,..,m-1. \tag{12}$$

Исследование устойчивости [6], проводимое путем возмущения правой части (2) некоторой величиной $\xi(x,t)$ и анализа отклика $v(x,t)$ при нулевых начальных условиях

$$\frac{\partial v}{\partial t}-\frac{\partial^2 v}{\partial x^2}=\xi(x,t), \xi(x,0)=0, \tag{13}$$

показывает, что справедливо неравенство

$$\max_{0\le j\le m}\max_{0\le i\le n}|v_{i,j+1}|\le T\max_{0\le j\le m}\max_{0\le i\le n}|\xi_{i,j+1}|, \tag{14}$$

причем для явной схемы это справедливо при $\lambda\le 1/2$. Отметим, что оценка (14) не учитывает погрешности округления.

Задачей настоящей работы является исследование, как убывающих регулярных компонент погрешности, так и составляющей $\Delta(n)$ при различных значениях независимых переменных x, t, n и m (или x, t, n и λ).

3. Результаты исследований

Для исследования величина n изменялась от значения $n{=}5$ до $n{=}5120$ последовательным удвоением текущего значения.

На рис. 1 в логарифмической шкале представлены результаты сравнения с точным решением данных, полученных при численном решении задачи с условиями вида I по явной схеме (12) для $\lambda{=}1/8$; $x{=}0.2$; $T{=}0.1$. По оси ординат отложены десятичные логарифмы (со знаком минус) относительных погрешностей, т.е. точность, выраженная в количестве точных десятичных значащих цифр. По оси абсцисс отложены десятичные логарифмы n. При таком выборе шкалы каждая компонента зависимости (1) отображается прямой линией.

Линии, обозначенные цифрой *0* на рис. 1, соответствуют точности вычисленных непосредственно значений разностной производной, цифрами *1,2* и т.д. – точности результатов первой, второй и т.д. фильтрации.

В результате каждой фильтрации меняется тангенс угла наклона линий (равный k_1, k_2, ... для $j{=}0, 1, \ldots$) и происходит сдвиг вверх, что свидетельствует о повышении точности. Такое поведение линий подтверждает наличие и результат устранения конкретных компонент зависимости.

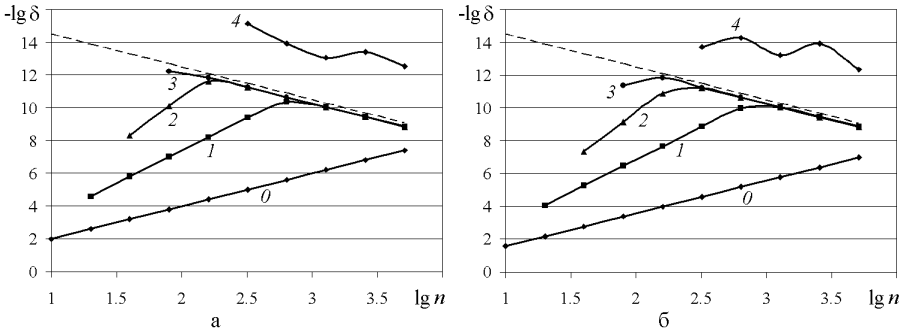

Рис. 1. Оценки погрешностей для явной схемы при $\lambda = 1/8$: а – искомых величин $u(x,t)$; б – производных $\frac{\partial u}{\partial x}(x,t)$. Пунктирная прямая y=16.5 − 2 · lg n

На рис. 1,а,б линии *1 – 3* имеют два четко выделенных участка. Первый участок имеет наклон, соответствующий показателю степенной функции. На втором участке угловой коэффициент приближенно равен -2. Это связано с преобладанием составляющей погрешности $\Delta(n)$, которую можно приближенно аппроксимировать прямой y=16.5 − 2 · lg n.

Анализ этих графиков позволяет установить, что составляющие погрешности метода $c_l n^{-k_l}$ имеют такие же значения показателей, как вторая разностная частная производная $\frac{\partial^2 u}{\partial x^2}(x,t) \approx \frac{u_{i+1,j}-2u_{i,j}+u_{i-1,j}}{h^2}$, т.е. k_l =2, 4, 6,... Отсутствие нечетных слагаемых, которые имеют место в разложении первой частной производной по времени, объясняется тем, что для сохранения значения λ при удвоении n, величину m приходится увеличивать в 4 раза, т.е. $\lambda m = n^2$.

Однако графики показывают два "нетипичных" свойства накапливающейся погрешности округления. Первое обнаруживается при анализе ее зависимости от n. Зависимость y=16.5 − 2 . lg n на рис. 1 является, в отличие от численного дифференцирования и интегрирования, не хаотичной, а близкой к закономерной. Эта компонента cn^2 может быть отфильтрована, и результат фильтрации, подтверждающий "закономерность" этой компоненты, показан на рис. 1 линиями с номером *4*.

Второе "нетипичное" свойство обнаруживается при сравнении линий на рис. 1,а и 1,б. При численном определении первой производной погрешность округления отсчитывается от исходного уровня нерегулярной погрешности как $\sim \frac{\Delta_0}{h}$, при этом $\frac{\Delta_0}{h} \approx \frac{cn^2}{h} = cn^3$. Однако вычислительный эксперимент обнаруживает погрешность на уровне cn^2. Тем самым, можно сделать вывод о том, что погрешность округления, имеющая порядок cn^2, накопленная при решении задачи с помощью разностной схемы (12), имеет регулярный характер в зависимости также и от переменной x.

На вопрос, является ли источником данной регулярной компоненты $\Delta(n)$ погрешность округления, помогает ответить применение попарного суммирования при вычислении по явной схеме. Из рис. 2,а следует, что в этом случае погрешность округления при увеличении n ограничена на уровне $10^{-15} - 10^{-16}$, т.е. отмеченный эффект полностью исчезает.

Результаты решения задачи с помощью неявной схемы с условиями II (6) при m=2n при применении метода итераций приведены на рис. 2,б. В отличие от рассмотренных выше результатов решения с помощью явной схемы, имеют место как четные, так и нечетные составляющие погрешности метода $c_l n^{-k_l}$. Это объясняется линейным ростом m (в связи с чем $\lambda = 0.5n^2/m \to \infty$) и первым порядком

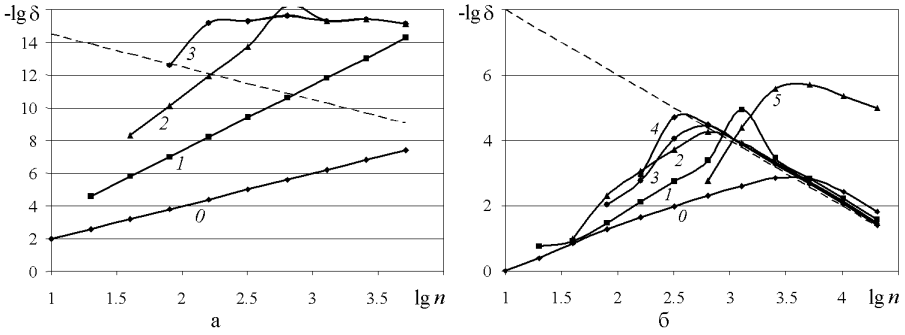

Рис. 2. Оценки погрешностей: а – для явной схемы с попарным суммированием; б – для неявной схемы с итерационным решением системы уравнений (11). Пунктирные прямые $y{=}16.5-2\cdot\lg\ n$ и $y{=}10-2\cdot\lg\ n$

аппроксимации первой частной производной по t. Что касается погрешности округления, то она имеет такие же свойства, как и в явной схеме (квадратично растет с увеличением n). Разница заключается в том, что погрешность решения системы уравнений итерационным методом ограничивалось значением $\varepsilon = 10^{-9}$. Это значение, а не длина мантиссы машинного слова (около 16-ти десятичных разрядов) определяло коэффициент зависимости cn^2 (хотя c совпадает с ε только приближенно, но имеет место приближенно прямо-пропорциональная зависимость от ε).

В связи с приведенными результатами представляет интерес исследование зависимости составляющей погрешности $\Delta(n)$ также и от x. На рис. 3 приведены зависимости погрешности округления (отнесенной к точному решению и к n^2) от x при разных n для неявной схемы: на рис. 3,а для условий вида I, на рис. 3,б для условий вида II (линии *1–4* соответствуют n=1280; 2560; 5120; 10240).

Из рис. 3,а,б видно, что относительная погрешность округления почти не зависит от x. Это само по себе не вызывает вопросов, так как погрешность округления определяется порядком числа в машинном представлении (в данном случае это $u_{i,j}$). Вызывает вопросы регулярность этой зависимости, причем как по x, так и по n.

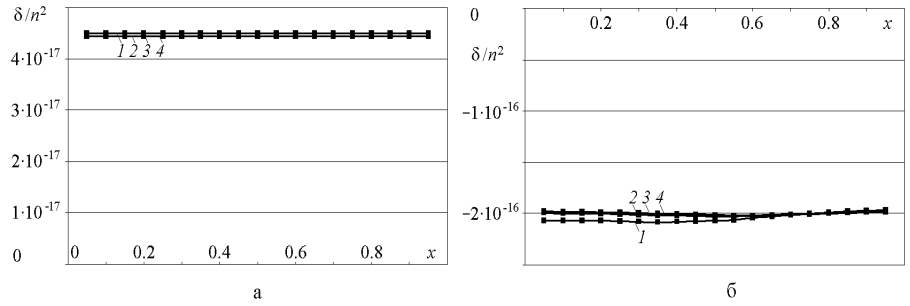

Рис. 3. Зависимости относительной погрешности округления от x при разных n для неявной схемы: а – с условиями вида I: б – с условиями вида II

Отметим, что при зависимости погрешности вида $\delta \approx cn^2$ отношение δ/n^2 приближенно определяет константу c. Представляет интерес исследовать зависимость $c(\lambda)$. На рис. 4 представлены зависимости относительной погрешности округления от λ для неявной схемы с условиями вида I. Графики построены с разными

шагами определения параметра λ: на рис. 4,б шаг в 4 раза меньше, чем на рис. 4,а. Видна "фрактальность" этой зависимости. Итак, по параметру λ зависимость имеет квазислучайный характер с ограничением $|c(\lambda)| \le 10^{-16}$. Для других видов условий это свойство также имеет место.

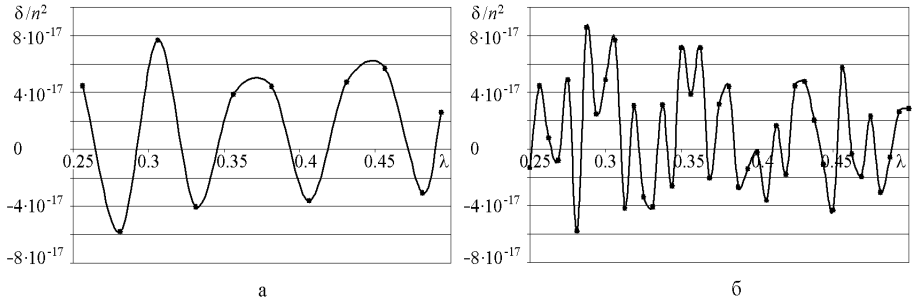

Рис. 4. Графики зависимости относительной погрешности округления от λ для неявной схемы, построенные с разными шагами определения параметра: а – по 10 точкам; б – по 40 точкам

Перейдем к исследованию зависимости погрешности округления от времени. На рис. 5,а приведены зависимости относительной погрешности округления (отнесенной к точному решению (8)) от времени для неявной схемы при λ=0.5 с шагом по безразмерному времени τ=0.05 (в диапазоне $0 \le t \le 1$) с условиями типа I. Погрешность округления найдена путем вычитания результатов, вычисленных с машинными словами с разной длиной мантиссы (16 и 19 десятичных разрядов).

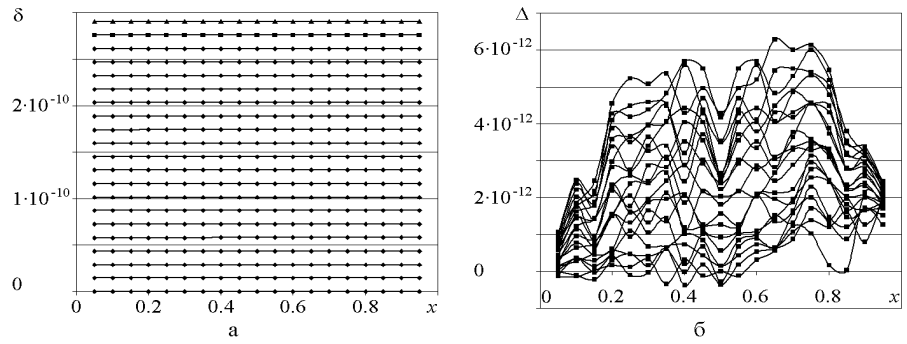

Рис. 5. Зависимости погрешности округления от времени при λ=0.5: а – с шагом τ=0.05; б – первые шаги при n=2560

Эксперимент показывает, наличие закономерной зависимости как от x, так и от t. Тем не менее, исследование первых шагов работы алгоритма (11) (результаты, полученные алгоритмом (12) имеют аналогичный вид) показывает, что закономерные зависимости складываются из квазислучайных элементов.

Для большей наглядности на рис. 6 приведены графики зависимостей погрешности округления от времени для x=0.05 (линии *1*) и x=0.5 (линии *2*). Видно, что абсолютная погрешность (рис. 6,а) сначала возрастает, а затем начинает убывать и приближается к нулю. При этом зависимость относительной погрешности от времени близка к линейной (рис. 6,б).

Такой же вид имеют зависимости от времени погрешности численного метода, только при этом порядок величин отличается, и зависимость от n имеет противоположный вид (погрешность округления растет, а погрешность метода уменьшается

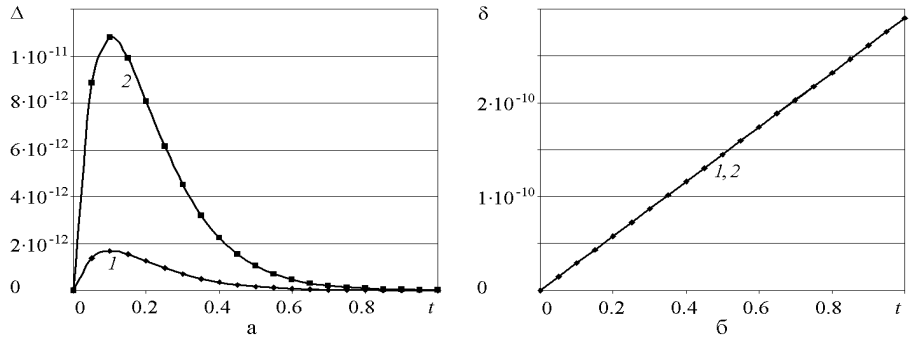

Рис. 6. Зависимости погрешности округления от времени при λ=0.5: а – абсолютной; б – относительной

при увеличении n).

Заметим, что теоретическая оценка сверху (14) показывает монотонно растущую с увеличением t зависимость, что в данном случае не имеет места. Этот факт опровергает наличие "чисто накопительного" механизма образования погрешности, лежащего в основе оценки (14).

4. Модели образования и изменения погрешности округления

Экспериментальный факт линейной зависимости относительной погрешности округления от времени позволяет предложить следующую модель образования и изменения погрешности округления. Поскольку величина погрешности округления определяется порядком операндов формул (11) и (12), рассмотрим задачу, аналогичную (13)

$$\frac{\partial v}{\partial t} - \frac{\partial^2 v}{\partial x^2} = c_1 u\left(x,t\right), v\left(0,t\right) = v\left(1,t\right) = 0, \tag{15}$$

или для условий вида III $\frac{\partial v}{\partial x}\left(0,t\right) = \frac{\partial v}{\partial x}\left(1,t\right) = 0$,
где $c_1 u\left(x,t\right)$ – модель возникающей при выполнении арифметических действий погрешности округления (возмущения); c_1 – некоторое число, не зависящее от x и t, которое связано с коэффициентом обнаруженной выше зависимости: $c_1 = cn^2$ ($|c| \le 10^{-16}$); $u\left(x,t\right)$ – решение задачи (2); v – отклик на возмущение правой части (15).

Для задачи с условиями вида I $u\left(x,t\right) = e^{-\pi^2 t}\sin\left(\pi x\right)$, тогда частное решение (15) $v\left(x,t\right) = t \cdot c_1 e^{-\pi^2 t}\sin\left(\pi x\right) = t \cdot u\left(x,t\right)$, что и показывает рис. 6,б.

Для задачи с условиями вида II точное решение дается формулой (9). Для простоты рассмотрим предельное решение $u\left(x,\infty\right) = 1 - x$. Частным решением (15) с данной правой частью является $v\left(x,t\right) = c_1\left[1 - x - \left(1-x\right)^3\right]\Big/6$. На рис. 7,а приведены зависимости погрешности округления, отнесенной к этому предельному решению. Видно, что при увеличении времени устанавливается величина слабо зависящая от x.

Аналогично, для задачи с условиями вида III точное решение дается формулой (10) с предельным решением $u\left(x,\infty\right) = 1/2$. Частным решением (15) с данной правой частью является $v\left(x,t\right) = c_1 t/2$. Зависимости погрешности округления, приведенные на рис. 7,б, аналогичны рис. 7,а.

На рис. 8 даны графики зависимостей относительной погрешности округления от времени для x=0.05 (линии *1*) и x=0.5 (линии *2*). Видно, что при условиях II погрешность возрастает и устанавливается на уровне $7 \cdot 10^{-11}$ (рис. 8,а), а при

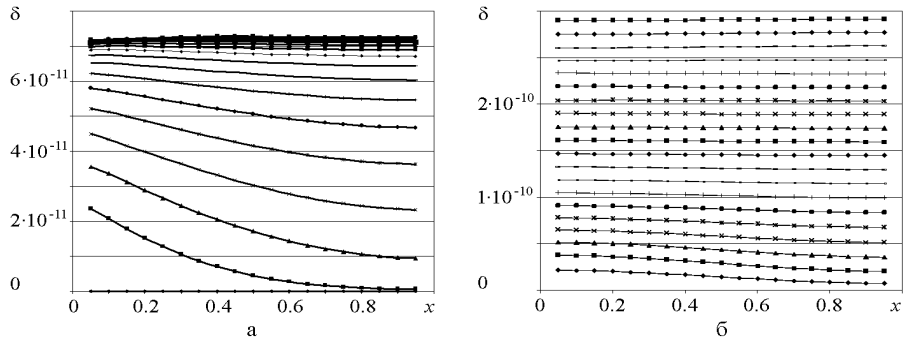

Рис. 7. Зависимости погрешности округления от времени при λ=0.5 с шагом τ=0.05: а – для условий типа II; б – для условий типа III

условиях III линейно растет, что согласуется с предложенной моделью (15). В последнем случае абсолютная и относительная погрешности отличаются только коэффициентом.

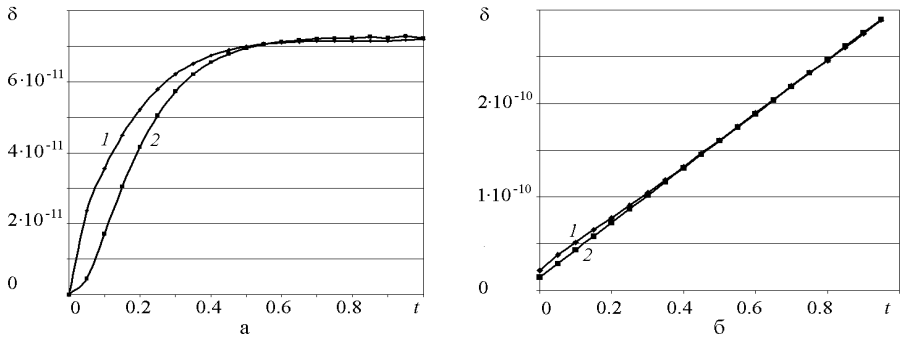

Рис. 8. Зависимости относительной погрешности округления от времени при λ=0.5, n=2560: а – для условий типа II; б – для условий типа III

Следует отметить, что условия III (второго рода) соответствуют изолированным концам, когда поток (жидкости, тепла) через границы отсутствует. При этом погрешность округления накапливается линейно, как предсказывает оценка (14). В остальных случаях устанавливается баланс, т.е. стационарное решение, когда количество принятой от источника сущности (т.е. жидкости или тепла) равно количеству ушедшей через границы. Интересно, что погрешность округления ведет себя так же, как жидкость или тепло.

На рис. 9 определяется коэффициент $c = c_1/n^2$ для вариантов условий I и III для разных n (кривые *1* – *7* соответствуют n=20; 40; 80; 160; 320; 640; 1280; 2560). Видно, что для $n \geq 160$ кривые близки к прямым и практически сливаются.

Отметим, что частные решения задачи (15), приведенные выше, являются аттракторами, так как общие решения (при различных начальных условиях) содержат кроме частного решения экспоненциально убывающие компоненты, что и показывает вычислительный эксперимент.

Результаты, представленные на рис. 3–9, получены с помощью неявной схемы при решении систем линейных алгебраических уравнений (СЛАУ) (11) методом прогонки.

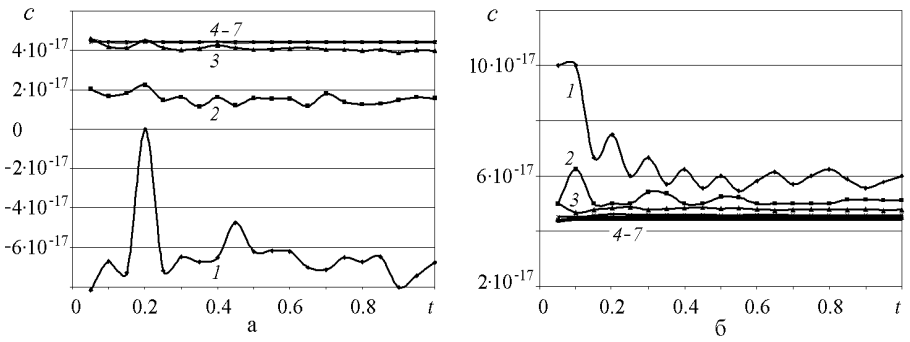

Рис. 9. Зависимости относительной погрешности округления от времени при λ=0.5: а – для условий типа I; б – для условий типа III

Рассмотрим результаты решения задачи (2) по неявной схеме с итерационным методом решения СЛАУ. В этом случае источником погрешности является не округление при выполнении арифметических операций, а погрешность решения СЛАУ, регулируемая сравнением максимальной по модулю разности последовательных итераций с заданным числом ε (см. рис. 2,б). В связи с этим правую часть дифференциального уравнения, моделирующего погрешность, необходимо изменить

$$\frac{\partial v}{\partial t} - \frac{\partial^2 v}{\partial x^2} = c_2 g\left(x\right), \tag{16}$$

где $g\left(x\right)$ – множитель, зависящий от x наиболее медленной из убывающих компонент разложений (8) – (10). Для задачи с условиями I и II $g\left(x\right) = \sin\left(\pi x\right)$, с условиями III $g\left(x\right) = \cos\left(\pi x\right)$. Для величины c_2 верхней оценкой является значение ε, т.е. $|c_2| \leq \varepsilon$. Это можно объяснить тем, что при шаге по времени при достаточно больших t разность между значениями искомого $u\left(x,t\right) - u\left(x,t+\tau\right) \approx c_3 e^{-\pi^2 t}\left(1 - e^{-\pi^2\tau}\right) g\left(x\right)$, что является начальной погрешностью в итерационном процессе. В процессе решения СЛАУ невязки уменьшаются, но распределение погрешности по x остается приблизительно пропорциональным $g\left(x\right)$. Итерационный процесс заканчивается, когда абсолютная величина разности, т.е. коэффициента перед $g\left(x\right)$ станет меньше ϵ. Частным решением такой задачи является $v\left(x,t\right) = \pi^{-2} c_2 g\left(x\right)$, что и подтверждается вычислительным экспериментом.

Необходимо сделать одно замечание относительно квадратичного роста погрешности округления от n. Оценка (14) получена путем перехода к разностным схемам уравнениям, возмущенным погрешностью $\xi\left(x,t\right)$

$$-\lambda v_{i-1,j} + \left(1+2\lambda\right) v_{i,j} - \lambda v_{i+1,j} = v_{i,j-1} + \tau\xi_{i,j}, \tag{17}$$

$$v_{i,j+1} = \lambda v_{i-1,j} + \left(1-2\lambda\right) v_{i,j} + \lambda v_{i+1,j} + \tau\xi_{i,j}, \lambda \leq 1/2. \tag{18}$$

Отсюда следует неравенство

$$\max_{0<i<n} |v_{i,j+1}| \leq \max_{0\leq i\leq n} |v_{i,j}| + \tau \max_{0<i<n} |\xi_{i,j}|, j = 0,..,m-1,$$

что и приводит к оценке (14).

Для оценки погрешности округления вместо (17) и (18) следует записать

$$-\lambda v_{i-1,j} + \left(1+2\lambda\right) v_{i,j} - \lambda v_{i+1,j} = v_{i,j-1} + \xi_{i,j},$$

$$v_{i,j+1} = \lambda v_{i-1,j} + \left(1-2\lambda\right) v_{i,j} + \lambda v_{i+1,j} + \xi_{i,j},$$

поскольку погрешность округления определяется порядком операндов $u_{i,j}$ и не зависит от τ. Но тогда изменится и оценка (14)

$$\max_{0\le j\le m}\max_{0\le i\le n}|v_{i,j+1}| \le \frac{T}{\tau}\max_{0\le j\le m}\max_{0\le i\le n}|\xi_{i,j+1}|\,, \tag{19}$$

и, если число Куранта λ сохраняется при увеличении n, то $\tau\sim n^{-2}$, что, вроде бы объясняет квадратичный рост погрешности. Однако это не объясняет сохранения зависимости погрешности от n при изменении связи τ с n для неявной схемы (рис. 2,б).

5. Особенности образования погрешности округления для явной схемы

Для анализа погрешности округления явная схема имеет преимущество по сравнению с неявной в связи с простотой проводимых арифметических операций. Для уменьшения числа операций умножения, приводящих к дополнительной погрешности, видоизменим схему (12)

$$u_{i,j+1} = u_{i,j} + \lambda\,(u_{i-1,j} - 2u_{i,j} + u_{i+1,j})\,. \tag{20}$$

Кроме того, вместо формулы $\lambda=\tau/h^2$ зададим непосредственно $\lambda=\frac{1}{k}$, $k=2,\ 3,\ 4,\ \ldots$ При T=0.1, $n=10,\ 20,\ 40,\ \ldots$, $h=1/n$, $\tau=T/m$ число $m=Tn^2/\lambda=kn^2/10$ является целым.

После таких видоизменений алгоритма эффект установления закономерной зависимости погрешности округления от n исчезает (рис. 10,а, кривые *1–5* соответствуют n=320, 640, 1280, 2560, 5120). При этом закономерность по x сохраняется. Однако при малом изменении λ на величину $\Delta_\lambda=-1\cdot10^{-16}$ эффект снова появляется при $c\approx1\cdot10^{-17}$ (рис. 10,б).

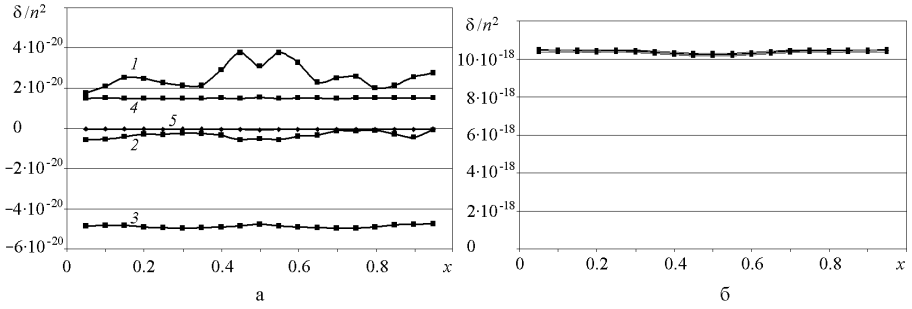

Рис. 10. Зависимости относительной погрешности округления от x при разных n для явной схемы с условиями вида I для $\lambda=1/2$: а – $\Delta_\lambda=0$; б – $\Delta_\lambda=-1\cdot10^{-16}$

При изменении знака приращения $\Delta_\lambda=1\cdot10^{-16}$ меняется знак c при сохранении модуля (рис. 11,а). Более того, при увеличении Δ_λ до величины порядка $1000\cdot10^{-16}$ значение $c\approx-1\cdot10^{-17}$ практически не меняется (рис. 11,б). При дальнейшем увеличении Δ_λ зависимость погрешности от n начинает “разваливаться”.

Тем самым, зависимость $c\,(\Delta_\lambda)$ представляет собой функцию типа $c_0 sign\,(\Delta_\lambda)$ при $c_0\approx1\cdot10^{-17}$. Аналогичное поведение имеет зависимость $c\,(\Delta_\lambda)$ при отсчете от других $\lambda=\frac{1}{k}$ при четных k (расчеты проводились до $k=20$). Более того, значение $c_0\approx-1\cdot10^{-17}$ сохраняется для всех четных k.

Но для нечетных k эффект установления квадратичной зависимости погрешности округления от n не возникает даже при больших Δ_λ.

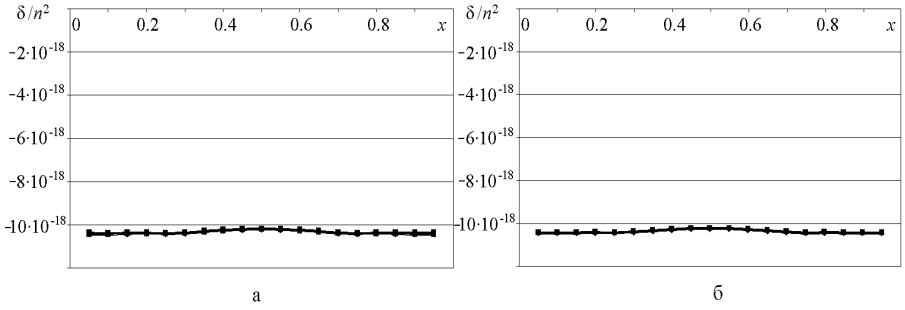

Рис. 11. Зависимости относительной погрешности округления от x при разных n для явной схемы с условиями вида I для $\lambda = 1/2$: а – $\Delta_\lambda = 1 \cdot 10^{-16}$: б – $\Delta_\lambda = 1000 \cdot 10^{-16}$

Все это справедливо для всех рассмотренных видов краевых условий. При этом меняется только c_0.

Проводились также исследования зависимости $c_0(\lambda)$ в ограниченных диапазонах с фиксированным шагом, при этом эффект то появлялся, то исчезал, значение $c_0(\lambda)$ иногда совпадало, иногда менялось, т.е. характер зависимости нельзя назвать ни чисто случайным, ни закономерным.

Следует отметить, что представленная выборка иллюстраций является лишь малой долей из множества проведенных экспериментов с помощью разных трансляторов и с использованием различных размеров машинных слов. На разных трансляторах даже одного языка программирования могут получаться количественно разные результаты с сохранением качественного совпадения. При этом для коротких (4-х байтных) слов при больших n возникает более грубая погрешность отсечения слагаемых, которые становятся по абсолютной величине меньше, чем единица последнего разряда второго слагаемого (20).

6. Выводы

Таким образом, с помощью вычислительного эксперимента выяснено, что зависимость погрешности численного метода от числа узлов n (1) состоит из тех же компонент, что и погрешность разностных производных. При этом абсолютное значение этой погрешности для вариантов условий I и II (краевые условия 1-го рода) при увеличении времени сначала растет, затем убывает и приближается к нулю. Отнесенная к функции (8) эта погрешность линейно возрастает. Для варианта III (краевые условия 2-го рода) справедливо то же самое, если отнести к функции, аналогичной (8) при замене синуса на косинус. Тем самым, модель источников (15) применима для исследования влияния на результат погрешности разностных производных.

Поведение погрешности округления в рассмотренной задаче намного сложнее, чем в задачах численного дифференцирования, интегрирования функций и решения задачи Коши для обыкновенных дифференциальных уравнений, в которых эта погрешность представлялась случайной величиной. В рассмотренной выше задаче на фоне случайной природы этой погрешности возникали закономерные зависимости от пространственной координаты, времени и числа узлов сетки n. На основе анализа результатов вычислительного эксперимента построена математическая модель возникновения и изменения во времени погрешности округления. В этой модели исходная погрешность, возникающая при выполнении арифметических действий, выступает в роли источника, распределенного в пространстве

пропорционально искомой величине. Зависимость коэффициента, определяющего общую интенсивность источника от числа Куранта λ, в неявной схеме представляется квазислучайной величиной, ограниченной единицей последнего разряда машинного слова. В явной схеме существует ряд значений λ, при которых эффекта возникновения регулярной зависимости от n не возникает, но при малых отклонениях от этих чисел она появляется. Причина возникновения или невозникновения явления формирования закономерности для явной схемы не объясняется экспериментом, и выяснение ее является задачей будущих исследований.

Благодарности

Работа выполнена при финансовой поддержке Минобрнауки России в рамках базовой части госзадания образовательным организациям высшего образования. Код программы 2229, год 2014.

Summary

V.P. Zhitnikov, N.M. Sherykhalina, S.S. Porechny. Peculiarities of the error accumulation process in solving the diffusion equation by finite difference methods.

A mixed problem for one-dimensional heat conductivity equation with a few variants of initial and boundary conditions is under consideration. The explicit and implicit schemes are applied for the solution. The sweep and the iteration methods are used for the equations system solving for the implicit scheme. Analysis of a calculating method error and a roundoff error is bases on the numerical filtration of a finite set of results, each corresponding to a different mesh, and a number of mesh knots n increases. Moreover, the results obtained for several mantissa lengths are compared for investigation of the roundoff error.

The results of computational experiment show, that the numerical method error is represented as a sum of a few power components cjn-k with integer degrees ($k = k_j > 0$). The roundoff error is accumulated along with the increase of mesh knots number as cn^2. This dependence on n has determinate character in contrast to the methods of numerical differentiation and integration. The coefficient c value either is bounded by the magnitude 10^{-M} (M is mantissa length) or a threshold error of the iteration method. The value of coefficient c changes chaotically with variation of the Courant number.

Key words: error of computation, passage from random to deterministic model, source model of error, rational Courant numbers.

Литература

1. *Житников В.П., Шерыхалина Н.М.* Моделирование течений весомой жидкости с применением методов многокомпонентного анализа. Уфа: Гилем, 2009. – 336 с.

2. *Житников В.П., Шерыхалина Н.М.* Многокомпонентный анализ численных результатов // LAP LAMBERT Academic Publishing, Germany, 2012. – 389 с.

3. *Житников В.П., Шерыхалина Н.М.* Применение многократной фильтрации при численном решении задач методами теории функций комплексного переменного // Вычислительные технологии. – 2013. – Т. 18. – № 1, – С. 15–24.

4. *Житников В.П., Муксимова Р.Р.* Решение задачи нестационарной электрохимической обработки плоским электрод-инструментом с ограниченной неровностью // Вестник УГАТУ. – 2011. – Т. 15. – № 1 (41). – С. 113–118.

5. *Борщук О.С., Житников В.П.* Нелинейная фильтрация в низкопроницаемых коллекторах. Численная схема, анализ устойчивости и сходимости // Научно-техн. Вестник ОАО "НК "Роснефть". – 2013. – № 2. – С. 13–16.

6. *Волков Е. А.* Численные методы. 2-е изд. испр. и доп. – М.: Наука, 1988. – 248 с.

Сведения о каждом из авторов статьи

Житников Владимир Павлович – д-р физ.-мат. наук, профессор, профессор Уфимского государственного авиационного технического университета

E-mail: *zhitnik@mail.ru*

Шерыхалина Наталия Михайловна – д-р техн. наук, доцент, профессор Уфимского государственного авиационного технического университета

E-mail: *n_sher@mail.ru*

Поречный Сергей Сергеевич – канд. физ.-мат. наук, доцент, доцент Уфимского государственного авиационного технического университета

E-mail: *porechny@mail.ru*